EXHIBIT A

# Molego-Based Definition of the Architecture and Specificity of Metal-Binding Sites

**Catherine H. Schein,[*] Bin Zhou, Numan Oezguen, Venkatarajan S. Mathura, and Werner Braun**
*Sealy Center for Structural Biology, Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, Galveston, Texas*

*ABSTRACT* **Decomposing proteins into "molegos," building blocks that are conserved in sequence and 3D-structure, can identify functional elements. To demonstrate the specificity of the decomposition method, the PCPMer program suite was used to numerically define physical chemical property motifs corresponding to the molegos that make up the metal-containing active sites of three distinct enzyme families, from the dimetallic phosphatases, DNase 1 related nucleases/phosphatases, and dioxygenases. All three superfamilies bind metal ions in a β-strand core region but differ in the number and type of ions needed for activity. The motifs were then used to automatically identify proteins in the ASTRAL40 database that contained similar motifs. The proteins with the highest PCPMer score in the database were primarily metal-binding enzymes that were related in function to those in the alignment used to generate the PCPMer motif lists. The proteins that contained motifs similar to the dioxygenases differed from those found with PCP-motifs for phosphatases and nucleases. Relatively few metal-binding enzymes were detected when the search was done with PCP-motifs defined for interleukin-1 related proteins, which have a β-strand core but do not bind metal ions. While the box architecture was constant in each superfamily, the specificity for the metal ion preferred for enzymatic activity is determined by the pattern of carbonyl, hydroxyl or imadazole groups in key positions in the molegos. These results have implications for the design of metal-binding enzymes, and illustrate the ability of the PCPMer approach to distinguish, at the sequence level, structural and functional elements. Proteins 2005;58:200–210. © 2004 Wiley-Liss, Inc.**

Key words: **PCPMer; MASIA; total sequence decomposition; DNase 1 superfamily; metal ion catalysis; dioxygenases; identifying functional homologues; dimetallic phosphatases; interleukin-1 structural family; protein design**

## INTRODUCTION

Metalloenzymes must maintain a delicate balance, binding ions tightly enough to retain them in the biological environment, while simultaneously allowing sufficient free sites for reactant binding.[1] The active sites of these enzymes contain a flexible network of carbonyl, hydroxyl, cysteinyl, and imadazole sidechains for inner shell coordination of metal ions, while still allowing interactions with the reactive groups of the substrates.[2,3] In previous work, we used a novel, word-based approach to parse aligned protein sequences of the APE1 family of nucleases, which is a subfamily of the DNase 1 superfamily, into discrete sequence motifs. We named the conserved 3D-structural areas of these motifs "molegos," for protein building blocks.[4,5] Molegos in our usage are shorter and more defined protein structure segments than the whole domains referred to elsewhere as molecular legos.[6,7] Here we show that these decomposition methods can be used to distinguish types of metal-binding enzymes in sequence databases.

The first step in our procedure is to decompose aligned sequences of proteins into physical chemical property (PCP)-based motifs[8] with our MOTIFMAKER program.[4] The motifs defined by MOTIFMAKER can be used by the MOTIFMINER program to scan databases to identify sequences with similar physical chemical properties. Structural data can then be used to determine which motifs correspond to structural elements that are highly conserved in other proteins in a family or superfamily, and are, thus, generally used molegos. In our previous work, we used this technique to determine the molegos that were common to both a non-specific nuclease (DNase 1) and a specific one, apurinic/apyrimidinic endonuclease (APE1) from those distinct for APE1. This allowed us to discriminate residues binding 3′ to the damage site in APE1, which were subsequently shown experimentally to be important

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Grant sponsor: Department of Energy; Grant number: DE-FG-00ER63041; Grant sponsor: John Sealy Memorial Endowment Fund; Grant number: 2535-01; Grant sponsor: Advanced Research Program of the Texas Higher Education Coordinating Board ATP (004952-0060-2001).

[*]Correspondence to: Catherine H. Schein, Sealy Center for Structural Biology, Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, Galveston, TX 77555-0857. E-mail: chschein@utmb.edu

Received 29 March 2004; Accepted 4 June 2004

Published online 25 October 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20253

for mediating substrate-binding specificity and proccessivity.[5,9]

In this report, we show that this approach can be used to distinguish homologues of metal-binding protein families in sequence databases. To explore how specifically metal-binding sites could be defined using our automated motif mining suite, PCPMer, we performed a decomposition analysis similar to that used for the DNase 1 family for two other well-studied families of metalloenzymes. The first are the di-metal ion centered phosphatases, which catalyze phosphorolytic cleavage of a variety of substrates. The second is the dioxygenases, a mono-metallic enzyme family involved in the oxidation of environmentally hazardous chemicals.[10] The dioxygenases are members of the (functionally extremely diverse) vicinal oxygen chelate (VOC) superfamily, which have similar metal-binding sites and common motifs, but bind different metal ions and substrates.[11] We defined PCP-motifs for the three enzyme families according to their physical chemical parameters, and used MOTIFMINER to scan the ASTRAL40 database to find proteins of known structure that contained similar sequences. This analysis revealed that the motifs in each case detected the enzymes in the initial alignment, and proteins with similar metal-binding properties and functions. That is, the proteins with the highest PCPmer scores, when the dioxygenase motifs were used to scan the database, were different from those found with the phosphatase motifs. Further, motifs from the interleukin-1 family of β-stranded growth factors, which are not known to bind metal ions, revealed many proteins that are related to this growth factor and receptor family but relatively few metal-binding proteins. This indicates that the combined PCPMer program, coupled with structural analysis, can serve as a useful aid for identifying distantly related homologues of a protein family.

## METHODS

### Physical Chemical Property Motifs and PCPMer

The PCPMer suite combines two programs, MOTIFMAKER and MOTIFMINER. The MOTIFMAKER program,[4] an outgrowth of our MASIA program,[12] searches for areas in aligned protein sequences that are conserved according to their physical chemical properties, based on a set of five vectors (E1–E5) that were defined by multidimensional scaling of 237 physicochemical properties of amino acid side chains.[8] The output of MOTIFMAKER is a series of numerical matrices for each motif in the protein that define the type and degree of conservation of the physical chemical properties of each column in the original sequence alignment. These matrices can then be used to automatically scan sequence databases, using the MOTIFMINER program, to identify proteins that contain sequences similar to the PCP-motifs defined for the initial set of proteins.[4]. Motifs can further be defined as "molegos," or molecular-building blocks, if their 3D-structure is conserved in the members of a family or superfamily where the motif occurs.

### Sequence Alignments

Motifs and molegos are defined for protein families that are recognizable homologues of one another. Alignments based on sequence data alone, using methods such as CLUSTALW, can be used if the sequences are not too diverse (preferably between 30 and 80% identical) and there are few gaps or insertions. For more diverse sequence families, such as those analyzed here, our previous work indicated that including structural information aids in properly aligning the sequences of known homologous proteins. Thus DALI[13] alignments of dimetallic phosphatases, dioxygenases, or interleukin-1 (IL-1) related proteins of known structure, were used as input to the MOTIFMAKER program (the original alignments and motif lists are given as supplementary data). We checked these alignments and the motifs generated by visual analysis of the structures and by using expert analysis of the families published by other groups.[10,11,14–16]

### Seqeunce Decomposition

Sequence decomposition of the APE1 family and analysis of related motifs in other members of the DNase 1 superfamily, using our MASIA tool (http://www.scsb.utmb.edu/masia/masia.html), was described previously.[5,12] PCP-motifs for the DNase 1 superfamily were isolated from an alignment of 17 diverse members of the DNase 1 superfamily (including 7 DNase 1 and 7 APEs from diverse species, and 3 IPPs of mammalian origin). PCP-motifs were extracted from the sequence alignments with the MOTIFMAKER subroutine of PCPMer (http://www.scsb.utmb.edu/PCPMer/),[4,8] using a specific entropy value of 1.25, allowed gap of 2, and a minimum length 5 (the alignment, the PCPMer motifs, and the scoring matrices for the motifs are given as supplementary data).

The 7 motifs that are common to the members of the DNase 1 superfamily are a subset of the 12 common to members of the APE subfamily.[4] To allow comparison with our previous report, the numbering of the molegos used in this study refers to the previously published list for the APE subfamily.[4,5] The APE1 motif 1, 2, 7, 11, and 12 correspond to the motifs 1, 2, 5–7 for the alignment of the DNase 1 superfamily.

Motifs and molegos of the dimetallic phosphatases were defined in MOTIFMAKER using a DALI alignment of 4 proteins of this superfamily of known structures. A sliding entropy definition was used and 18 motifs were defined. Motifs were defined similarly for a DALI alignment of three dioxygenase proteins that included the three metal-binding regions known to be similar in this family. Finally, a previously defined alignment of IL-1β homologues, all of which contain a similar β-stranded core,[17] was used as a non-metal-binding control for the PCPMer method.

### Database Searching

The MOTIFMINER subroutine of PCPMer was then used to score proteins in the ASTRAL40 database[18,19] (versions 55 and 63) according to their similarities to the PCP-motifs defined for the starting alignment. The ASTRAL40 database contains ~3,700 sequences of proteins, representing nearly every unique protein structure in the PDB. Protein scores can be derived in two ways, depending on the method chosen to determine a significant match. Where conservation is high, a cutoff value for significance can be specified (such as 0.7). Alternatively, a mean scoring system can be selected, to use the average score of the sequences in the starting alignment and that of all sequence windows in the database to determine a significance threshold.

Molego pictures were drawn with MOLMOL[20] from the indicated PDB files.

## RESULTS

### Molego Architecture of Three Metal-Binding Protein Families

Figure 1 shows representatives of the metal-containing active sites in the three enzyme superfamilies compared in this study, for the DNase 1 superfamily, the dimetallic phosphatases, and the dioxygenases. The molegos, in this case the conserved β-strands that make up the three sites, differ considerably between the three types of metalloenzymes in their topology and the relative location of the metal ion(s). One representative structure is shown for each of the three metalloenzyme groups discussed in this report. The first structure, for human APE1, represents the DNase 1 topology. We previously observed that the β-strand core of all enzymes of the DNase 1 superfamily is highly conserved[5] and particularly the five molegos that form the antiparallel active center of the enzymes.[4] For example, this area in inositol 5′ polyphosphate phosphatase, synaptojanin, a distantly related member of the DNase 1 fold family, is similar in sequence and geometry with that of APE1. The second molego drawing, for 5′ nucleotide phosphatase, represents the dimetallic phosphatase family, which contains two metal ions in the active site. Again, five β-strands make up its active center, but the overall topology is distinct from the monometallic APE1 site. The third β-core, for 2, 3-dihydroxybiphenyl 1, 2-dioxygenase, differs considerably from the other two structures in that the metal ion is bound in the middle of the β-strands, not at the ends.

Scanning of the proteins in the ASTRAL40 database with MOTIFMINER[4,8] revealed several metalloenzymes that contained sequence elements similar to the PCP-motifs of the DNase 1 superfamily (Table I and Mathura et al.[4]). Among these were many nucleases, RNA and nucleotide-binding proteins, and proteins with metal-binding capability. To determine the selectivity of the PCPMer approach, we did a similar structural decomposition and database search for the two other metal ion–binding superfamilies, and (as control), for the IL-1 family of proteins that have a similar β-stranded core but no known metal-binding capability.

### PCP-Motifs of the Dimetallic Phosphatases Detect Other Phosphatases

PCP-motifs for dimetallic phosphatases were identified by the MOTIFMAKER program in a DALI alignment of the sequences of four dimetallic phosphatases of known structure. The PCP-motifs were checked by comparing them to previously identified sequence motifs for one of the sequences, a representative of the nucleotide 5′ phosphatase family of proteins.[14] The molegos in the metal boxes of these enzymes are conserved across the superfamily, which includes enzymes with such diverse function as the DNA repair enzyme MreII (PDB 1ii7; CSOP d.159.1.4), pig acid phosphatase (1ute, SCOP d.159.1.1), and λ-phage serine/threonine protein phosphatase (1g5b, SCOP d.159.1.3) (Schein et al., forthcoming). The PCP-motifs that were defined for the phosphatases with MOTIFMAKER were then used to scan the ASTRAL40 database for sequences containing similar regions. MOTIFMINER results (Table II) show that the highest scoring proteins were the dimetallic phosphatases in the initial alignment, as well as a closely related protein phosphatase that was not included in that alignment. The other high-scoring proteins in this search were metalloenzymes that were similar in function to those of the starting alignment, and different from those found with the DNase 1 superfamily motifs (Table I).

### The Dioxygenases Have a Different Metal Ion Catalytic Center

To further determine the specificity of the PCPMer methodology, we decomposed the aligned sequences of a family of metalloenzymes that are not functionally related to the DNase 1 or the dimetallic phosphatase superfamilies. We chose the dioxygenases, a family within the vicinal oxygen chelate (VOC) superfamily of metalloenzymes that catalyze oxidative cleavage of C-C bonds, isomerizations, epimerizations, and nucleophilic substitutions. The motifs that characterize this superfamily have been shown to form a βαβββ structural unit in the metal-containing active center.[11] Compared to the phosphatases and nucleases, there are fewer protein ligands to the metal ion in the VOCs, presumably to allow tighter coordination between the substrate and the metal ion during the formation of the enolic intermediate.[16] The isolated molegos (Fig. 2) show how the β-strands of the dioxygenase metal site are conserved, regardless of the metal bound. Table III compares the sequence conservation of these three elements. While the first molego sequence is more variable, the other two are well conserved according to their physical chemical properties. The highlighted amino acids, made clear by the molego-based alignment in Table III, also illustrate how a small change (H to E in motif 1) may indicate selectivity for $Zn^{2+}$ in 1QIP. However, the pattern of change in the amino acids is not yet fully quantified, as will be discussed below for the ensemble of proteins, and will require observing the coordination spheres of more proteins in this family.

The sequences of the three elements were defined as PCP-motifs using MOTIFMAKER and these were used to scan the ASTRAL40 database. MOTIFMINER rapidly identified the three proteins in the initial alignment within the first top 20 proteins (Table IV). The intervening proteins with similar PCPMer scores were pre-

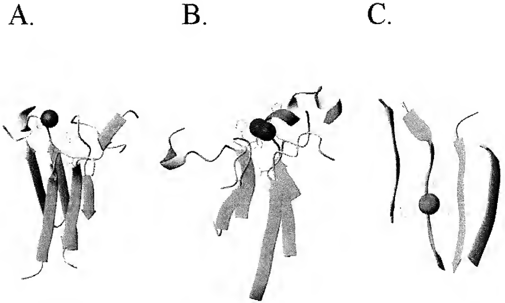

Figure 1.

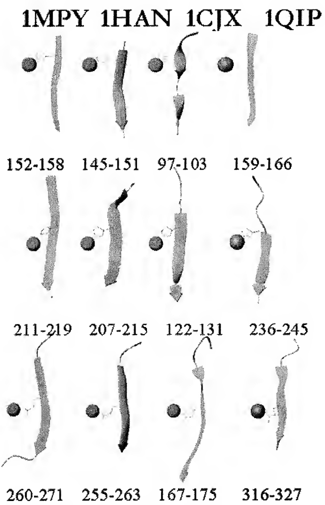

Figure 2.

dominantly metal binding, and included many oxidases. The list of metal-binding proteins identified starting from the dioxygenases was distinct from those found by scanning the database with the conserved PCP-motifs of the DNase 1 (Table I) and dimetallic phosphatases (Table II).

### Using PCP-Motifs to Identify β-Strand Proteins That Are Not Metalloenzymes

About one third of all proteins bind metal,[21] and novel metal-binding sites have also been found by structure analysis.[22] Many, but not all metal-binding sites[23,24] in metalloenzymes are composed of β-strands. To determine whether PCPMer was only recognizing the sequence patterns for β-strand formation and not metallo-binding sites, we isolated PCP-motifs from a structure-based alignment of proteins related to interleukin-1 (IL-1).[17] These proteins all have a β-strand core but are not known to bind metal ions. The highest scoring proteins related to this family in

Fig. 1. Molego representations of the metal-containing active site regions of three different metalloenzyme families. **A:** From the structure of human APE1 with Mn$^{2+}$ (PDB file 1DE9), a representative of the DNase1 related nucleases and phosphatases; B: 5′ nucleotide phosphatase with two Zn$^{2-}$ (PDB file 1USH), a representative of the dimetallic phosphatases; **C:** 2,3-dihydroxybiphenyl 1,2-dioxygenase with FeII (PDB file 1HAN), a representative of the dioxygenase family. The molego segments are shown in ribbon format (corresponding to their conserved secondary structures across a family or superfamily), including the side chains of key residues near the metal ions.

Fig. 2. Metal-binding molegos in three FeII binding dioxygenases (1MPY:catechol 2,3-dioxygenase;1HAN: 2,3-dihydroxybiphenyl 1,2-dioxygenase; 1CJX:4-hydroxyphenylpyruvate dioxygenase) and another member of the vicinal oxygen chelate superfamily (VOC) that binds zinc (1QIP, human glyoxylase). Note the metal ion binding residues are within the β-strands, rather than projecting above them, and that the structure is constant while the residues that bind the metal ion dictate the specificity.

**TABLE I. Highest Scoring Proteins in the ASTRAL40 Database of Representative PDB Files Selected by PCPMer, Using the Motif Profile of the DNase 1 Superfamily[†]**

| PCPMer score[a] | PDB ID | SCOP | EC number | Bound ion | Description |
|---|---|---|---|---|---|
| **1683** | **2DNJ** | **d.151.1.1[b]** | **3.1.21.1** | **$Mg^{2+}$** | **Deoxyribonuclease I {Cow (Bos taurus)}[c]** |
| **1604** | **1AKO** | **d.151.1.1[b]** | **3.1.11.2** | **$Mg^{2+}$, $Mn^{2+}$** | **DNA-repair enzyme exonuclease III (*E. coli*)[c]** |
| **1501** | **1HD7** | **d.151.1.1[b]** | **4.2.99.18** | **$Mg^{2+}$, $Mn^{2+}$** | **DNA repair endonuclease Hap1 (Hu)[c]** |
| **1472** | **1I9Z** | **d.151.1.2[b]** | **hydrolase** | **$Ca^{2+}$** | **Synaptojanin, IPP5C domain {Yeast (*Schizc S. pombe*)}[c]** |
| 1448 | 1QBK | a.118.1.1 | Nuc. trans | $Mg^{2+}$ | Karyopherin β2; nuclear transporter {Hu} |
| 1371 | 2BCE | c.69.1.1 | 3.1.1.13 | taurocholate | Bile-salt activated lipase (cholesterol esterase) {Cow (Bos taurus)} |
| 1365 | 1QQQ | d.117.1.1 | 2.1.1.45 | Nucleotide | Thymidylate synthase {*E. coli*} |
| 1364 | 1GQI | c.1.8.10 | 3.2.1.139 | $Co^{2-}$, $Mg^{2+}$ | (A: 152–712) alpha-D-glucuronidase catalytic domain {*Pseudomonas cellulosa*} |
| 1355 | 1E4M | c.1.8.4 | 3.2.3.1 | $Zn^{2-}$ | Plant beta-glucosidase (myrosinase) {White mustard (Sinapis alba)} |
| 1352 | 1F8M | c.1.12.6 | 4.1.3.1 | $Mg^{2+}$ | Isocitrate lyase {*Mycobacterium tuberculosis*} |
| 1350 | 1GPI | b.29.1.10 | 3.2.1.91 |  | Cellobiohydrolase I {Ce17d} |
| 1340 | 1I50 | e.29.1.1 | 2.7.7.6 | $Mg^{2+}$, $Ca^{2+}$ | RBP1 (*S. cerevisiae*) |
| 1339 | 1HO8 | a.118.1.9 | 3.6.1.34 | $Eu^{2-}$[d] | Regulatory subunit H of the V-type ATPase {Baker's yeast (*S. cerevisiae*)} |
| 1334 | 3BTA | d.92.1.7 | 3.4.24.69 | $Zn^{2-}$ | (A:1–546) Botulinum neurotoxin {*Clostridium botulinum*, serotype A} |
| 1319 | 1C8D | b.10.1.4 | Viral protein | $Ca^{2-}$ | Parvovirus (panleukopenia virus) capsid {Dog (*Canis familiaris*)} |
| 1315 | 1FBN | c.66.1.3 | Ribosome | RNA | Fibrillarin homologue {*Archaeon Methanococcus jannaschii*} |
| 1300 | 1QFX | c.60.1.3 | 3.1.3.8 | $PO_4^{3-}$ | Phytase (myo-inositol-hexakisphosphate-3-phosphohydrolase) {*Aspergillus niger*} |
| 1289 | 1M1X | b.69.8.1 | Nuc. transporter | $Mn^{2+}$ | (A:1–438) Integrin alpha N-terminal domain {Hu} |
| 1288 | 1KO6 | b.119.1.1 | Transferase | RNA binding | C-terminal autoproteolytic domain of nucleoporin nup98 {Hu} |
| 1287 | 1CLC | a.102.1.2 | 3.2.1.4 | $Ca^{2-}$, $Zn^{2+}$ | (135–575) CelD cellulase, C-terminal domain {*Clostridium thermocellum*} |
| 1282 | 1D1Q | c.44.1.1 | 3.1.3.48 |  | Tyrosine phosphatase {Baker's yeast (*S. cerevisiae*)} |
| 1279 | 1QAZ | a.102.3.1 | 3.5.1.45 | $SO_4^{2-}$ | Alginate lyase A1-III {*Sphingomonas sp.*, A1} |
| 1271 | 1QQ9 | c.56.5.4 | 3.4.11 | $Ca^{2-}$, $Zn^{2+}$ | Aminopeptidase {*Streptomyces griseus*} |
| 1268 | 1QQ1 | b.80.1.6 | Viral protein |  | P22 tailspike protein {*Salmonella phage*} |
| 1266 | 1A2V | b.30.2.1 | 1.4.3.6 | $Cu^{2+}$ | (A:237–672) Copper amine oxidase, domain 3 (catalytic) (*Hansenula polymorpha*) |
| 1254 | 1FIU | c.52.1.10 | 3.1.21.4 | $Mg^{2+}$ | Restriction endonuclease NgoIV {*Neisseria gonorrhoeae*} |
| 1248 | 1AYX | a.102.1.1 | 3.2.1.3 |  | Glucoamylase-{*Saccharomycopsis fibuligera*} |
| 1247 | 1BHE | b.80.1.3 | 3.2.1.15 |  | Polygalacturonase-*Erwinia carotovora* |
| 1238 | 1BVY | c.23.5.1 | 1.14.4.1 | Heme | FMN-binding domain of the cytochrome P450bm-3 {*Bacillus megaterium*} |
| 1238 | 1M1N | c.92.2.3 | 1.18.6.1 | $Fe^{2+}$ $MoO_4^{2-}$ | Nitrogenase iron-molybdenum protein, alpha chain {*Azotobacter vinelandii*} |
| 1237 | 1FN9 | d.196.1.1 | Viral protein | $Zn^{2-}$, dsRNA | Outer capsid protein sigma 3 {Reovirus} |
| 1237 | 1N1T | b.68.1.1 | 3.2.1.18 |  | (A:1–406) *Trypanosoma rangeli sialidase* |
| 1234 | 1F46 | d.129.4.1 | Cell cycle |  | Cell-division protein ZipA, C-terminal domain {*E. coli*} |
| 1221 | 1USH | d.159.1.2 | 3.1.3.5 | $Zn^{2-}$ | (26–362) 5′-nucleotidase (syn. UDP-sugar hydrolase), N-terminal domain {*E. coli*} |
| 1216 | 1CJA | d.144.1.3 | Transferase | AMP binding | Actin-fragmin kinase, catalytic domain {Slime mold (*Physarum polycephalum*)} |
| 1208 | 2SHP | c.45.1.2 | 3.1.3.48 | $PO_4^{3-}$ | (A:219–525 Tyrosine phosphatase {Hu, shp-2} |

[†] Motif profile was generated by PCPMer with a relative entropy of 1.25, a gap of 2, and a minimum length of 5.

[a] PCPMer uses a Bayesian scoring function to determine proteins that contain the highest scoring matching motifs.

[b] The SCOP class, d.151.1, is for the DNase 1 superfamily.

[c] Sequences in the initial alignment are **bold.**

[d] Europium ions used to obtain phase data may demarcate calcium binding sites.

**TABLE II. The Proteins in the ASTRAL40 Database (Version 1.63) That Most Closely Match the PCP-Motifs of the Dimetallic Phosphatases Are Predominantly Metal-Binding Proteins†**

| PCPMer score | PDB code | SCOP | EC number | Bound ion | Description |
|---|---|---|---|---|---|
| **5519** | **1UTE** | **d.159.1.1** | **3.1.3.2** | **Fe$_2$O$^{4+}$** | **Purple acid phosphatase {Pig}** |
| **5420** | **1USH** | **d.159.1.2** | 3.1.3.5 | **Zn$^{2+}$, SO$_4^{2-}$, CO$_3^{2-}$** | **5′-nucleotidase (syn. UDP-sugar hydrolase), N-terminal domain {*E. coli*} (26–362)** |
| 5284 | 1AUI | d.159.1.3 | 3.1.3.16 | Ca$^{2+}$, Fe$^{3+}$, Zn$^{2+}$ | Ser/thr phosphatase-2B (PP-2B, calcineurin A subunit) {Hu} |
| **5283** | **1II7** | **d.159.1.4** | **Replication** | **Mn$^{2+}$, PO$_4^{3-}$, SO$_4^{2-}$** | **Mre11 {*Archaeon Pyrococcus furiosus*}** |
| 5279 | 1M7S | e.5.1.1 | 1.11.1.6 | Heme | Catalase I {*Pseudomonas syringae*} |
| **5217** | **1G5B** | **d.159.1.3** | **3.1.3** | **Hg$^{2+}$, Mn$^{2+}$, SO$_4^2$** | **ser/thr protein phosphatase {Bacteriophage λ}** |
| 5163 | 1LI5 | c.26.1.1 | 6.1.1.1 | Zn$^{2-}$ | Cysteinyl-tRNA synthetase (A:1–315) {*E. coli*} |
| 5150 | 2BCE | c.69.1.1 | 3.1.1.13 | Taurocholate | Bile-salt activated lipase (cholesterol esterase) {Cow (Bos taurus)} |
| 5139 | 1IAT | c.80.1.2 | 5.3.1.9 | SO$_4^{2-}$ | Phosphoglucose isomerase, PGI {Hu} |
| 5130 | 1RKD | c.72.1.1 | 2.7.1.15 | ADP, PO$_4^{3-}$ | Ribokinase {*E. coli*} |
| 5125 | 1I50 | e.29.1.2 | 2.7.7.6 | Zn$^{2-}$, Mn$^{2-}$ | RNA Polymerase II {*S. cerevisiae*} |
| 5122 | 1F6D | c.87.1.3 | 5.1.3.14 | Na$^+$, Cl$^-$, UDP | UDP-N-acetylglucosamine 2-epimerase {*E. coli*} |
| 5117 | 1MJG | e.26.1.3 | 1.2.99.2 | Cu$^{1+}$, Ni$^{2+}$, Fe$_4$S$_4^{2+}$ | Bifunctional carbon monoxide dehydrogenase/acethyl-CoA synthase α-subunit {*Moorella thermoacetica*} |
| 5113 | 1AOZ | b.6.1.3 | 1.10.3.3 | Cu$^{2+}$, Cu-O-Cu | Ascorbate oxidase {Zucchini; A: 339–552} |
| 5099 | 1IW7 | e.29.1.2 | 2.7.7.6 | Mg$^{2+}$, Pb$^{2+}$ | RNA-polymerase beta-prime {*Thermus thermophilus*} |
| 5095 | 1EHK | f.24.1.1 | 1.9.3.1 | Dinuclear Cu, heme | Bacterial ba3 type cytochrome c oxidase subunit I {*Thermus thermophilus*} |
| 5094 | 1JBO | f.29.1.1 | Photosynthesis | Ca$^{2+}$, Fe$_4$S$_4^{2-}$ | Apoprotein a1, PsaA {*Synechococcus elongatus*} |
| 5091 | 1PRE | f.8.1.1 | Toxin | — | (Pro) aerolysin, {Aeromonas hydrophila} (85–470) |
| 5083 | 1FLG | b.70.1.1 | Oxidore ductase | Ca$^{2+}$, PQQ | Ethanol dehydrogenase {*Pseudomonas aeruginosa*} |
| 5074 | 1QLW | c.69.1.15 | Hydrolase | SO$_4^{2-}$ | Bacterial esterase {*Alcaligenes sp.*} |
| 5071 | 1PBG | c.1.8.4 | 3.2.1.85 | SO$_4^{2-}$ | 6-phospho-beta-D-galactosidase, PGAL {*Lactococcus lactis*} |
| 5071 | 1G8K | c.81.1.1 | Oxidore ductase | Mo$^{4+}$, FeS, Hg$^{2+}$, Ca$^{2+}$ | Arsenite oxidase large subunit {*Alcaligenes faecalis*} (A:4–682) |
| 5070 | 1A9X | c.23.16.1 | Amidotransferase | K$^+$, Cl$^-$, Mn$^{2+}$, PO$_4^{3-}$ | Carbamoyl phosphate synthetase, small subunit C-terminal domain {*E. coli*} (B:1653–1880) |

†The first part of the PCPMer results file for proteins in the ASTRAL40 database (version 1.63) is shown. A DALI alignment of 4 dimetallic phosphatases (highlighted in **bold**) was used to define 15 motifs in MOTIFMAKER using a sliding relative entropy scale [range (0.5-1.7) step 0.1, gap cutoff 1, length cutoff 6]. The database search was done in MOTIFMINER, with the matching sequences scored with a cutoff value of 0.7 (i.e., only sequences with a score of 0.7 or higher to a given motif would be considered a match).

**TABLE III. Sequences of the Fe(II) Binding Motifs (See the corresponding molegos in Fig. 2) of the Dioxygenase Family (1MPY, 1HAN, 1CJX)†**

| PDB ID | Motif 1 | Motif 2 | Motif 3 |
|---|---|---|---|
| 1MPY | 152 DHALMYG 158 | 211 R-LHHVSFHL 219 | 260 SGNRNEVFCGG271 |
| 1HAN | 145 GHFVRCV 151 | 207 R-IHHFMLEV 215 | 255 SGVEVEYGW--263 |
| 1CJX | 159 DHLTHNV166 | 236 EGIQHVAFLT 245 | 316 GDVFFEFIQRK327 |
| 1QIP | 97 LELTHNW 103 | 122 RGFGHIGIAV 131 | 167 DGYWIEILN--175 |

†The corresponding area of another member of the vicinal oxygen chelate family, 1QIP, a lyase with a different metal binding specificity (Zn) is shown for comparison. Residues that are in direct contact with the metal ion are **bold.**

the ASTRAL40 database (Table V) were those in the initial alignment, cell surface receptors and viral structural proteins. Of the top 30 scoring proteins, only 8 were metalloenzymes, and several others contained metal ions that functioned in lattice formation. In comparison, for the phosphatase analysis, 17 of the top 24 sequences were metalloenzymes (Table II), as were 14 of the 23 highest scoring proteins for the dioxygenase motifs (Table IV). Thus, the program can distinguish different functional types of enzymes, and not just secondary structure regions.

## Correlating the Key Amino Acid(s) Presented by the Molego with Metal Ion Choice

One reason that MOTIFMINER is able to identify metal-binding motifs is that it uses not just the average value of residues in a column of the original alignment, but also the "relative entropy," a measure of the residue variability, in scoring motifs in the database sequences. The most conserved amino acids in the motifs described here are indeed those in direct contact with the metal ion, and

**TABLE IV. Proteins Identified in the ASTRAL40 Database That Contain Regions With Significant Similarity to the PCP Motifs That Define the Metal-Binding Site of Dioxygenases**

| PCPMer score | PDB code | SCOP | EC number | Bound ion | Description |
|---|---|---|---|---|---|
| 142.06 | 1FO3 | a.102.2.1 | 3.2.1.24 | $Ca^{2+}$ | HuClassIα-1; 2-mannosidase, catalyticdomain |
| 140.77 | 1LVK | c.37.1.9 | Contractile protein | $Mg^{2-}$ | MyosinS1, motor domain {(*Dictyostelium discoideum*)} |
| **140.28** | **1HAN** | **d.32.1.3** | **1.13.11.39** | $\mathbf{Fe^{2+}}$ | **2,3-Dihydroxybiphenyl dioxygenase**[a] |
| **140.28** | **1CJX** | **d.32.1.3** | **1.13.11.27** | $\mathbf{Fe^{2+}}$ | **4-hydroxyphenylpyruvate dioxygenase (*Pseudomonas fluorescens*)**[a] |
| 139.33 | 1GPR | b.84.3.1 | 2.7.1.69 | — | Glucose permease Iia domain, 11a-glc (*Bacillus subtilis*) |
| 138.66 | 1CJY | c.75.1.1 | 3.1.1.4 | $Zn^{2-}$ | Hu cytosolic phospholipase A2 |
| 137.86 | 1AVA | c.1.8.1 | 3.2.1.1 | $Ca^{2+}$ | Plant alpha-amylase {Barley} |
| 137.86 | 1CR1 | c.37.1.11 | 2.7.7 | Sulfate | g4p, DNAprimase, helicase domain {Bacteriophage T7} |
| 137.57 | 1BSX | a.123.1.1 | Hormone | Triiodothiamine | Hu Thyroid hormone receptor beta |
| 137.50 | 1CYD | c.2.1.2 | 1.1.1.184 | NADPH | Carbonyl reductase {Mouse} |
| 136.80 | 1DF0 | d.3.1.3 | 3.4.22.17 | $Ca^{2+}$ | Calpain (calcium dependent protease) {Rat} |
| 136.60 | 1QQT | c.26.1.1 | 6.1.1.10 | $Zn^{2+}$ | Methionyl-tRNAsynthetase |
| 136.54 | 1UAA | c.37.1.13 | 3.6.1 | ADP | DEXX box DNA helicase {*E. coli*} |
| 136.52 | 1E5M | c.95.1.1 | 2.3.1.41 | Acyl group | β-ketoacyl-ACP synthaseII {*Synechocystis sp.*} |
| 136.31 | 1VNS | a.111.1.3 | 1.11.1.10 | Vanadate | Chloroperoxidase {*Curvularia inaequalis*} |
| 136.26 | 1FL1 | b.57.1.1 | Viral protein | $K^-$ | Protease {Kaposi's sarcoma-associated herpes virus} |
| 136.05 | 1ELU | c.67.1.3 | lyase | $K^-$ | Cystine C-Slyase {Synechocystissp.}, Fe S assembly |
| 135.87 | 1QBK | a.118.1.1 | Nuclear transport | GNP, SeM, $Mg^{2+}$ | Hu-Karyopherin beta2 nuclear transporter |
| 135.87 | 1ZPD | c.36.1.1 | 4.1.1.1 | $Mg^{2-}$, | Pyruvate decarboxylase {*Zymomonas mobilis*} |
| 135.69 | 1FUR | a.127.1.1 | 4.2.1.2 | Malate | Fumarase {*E. coli*} |
| 135.23 | 1IHP | c.60.1.3 | 3.1.3.8 | Sulfate | Phytase (myo-inositol-hexakisphosphate-3-phosphohydrolase) |
| **135.15** | **1MPY** | **d.32.1.3** | **1.13.11.2** | $\mathbf{Fe^{2+}}$ | **Catechol2,3-dioxygenase {*Pseudomonas putida*}** |

[a]The C-terminus of the proteins in **bold** were in the DALI alignment used (MOTIFMAKER subprogram, PCPMer) to define the matrices for the motifs.

MOTIFMINER will give the highest scores to sequences that match at these positions. Comparison of the protein-binding molegos from the three superfamilies revealed that similar binding sites bound different metals, and that the key amino acids that dictated the metal ion choice were indeed the most conserved.

Table VI summarizes the metal-binding site and distances to nearby residues (within 3Å of the metal ion except for the DNase 1 family representatives, where the metal is more loosely bound) of all the enzymes in this study as a function of the preferred metal ion for catalysis (which is not always identical with that used for the crystal structure determination). The metal ions in several of the structures have bonds to substrates and water molecules, which for the sake of clarity have not been included in Table VI. For example, in the 1DE9 structure of HuAPE1 with $Mn^{2+}$, which is tetrahedrally coordinated, there are additional bonds to oxygen atoms in the substrate DNA. The $Ca^{2+}$ ion in the synaptojanin (1I9Y) structure has 6 ligands within a 3.5 Å radius, of which only two are from the protein. The rest are water ions. In 1QIP, the single $Zn^{2+}$ atom is coordinated by four protein ligands, and is also very close to the two oxygen atoms in the $O_2$ molecule in the active site. A full summary of the bonds to each metal ion in the PDB structures is included in Table VII, which is given as supplementary information.

All the metals are bound tightly by at least one carboxylic oxygen, from an aspartate or a glutamate. The other residues in the binding site differ in a fashion that indicates their metal ion specificity, but the exact pattern must be determined by a more complete comparison of the molegos in other members of the families. As expected from previous analysis of hydration[25] and bonding patterns of metal ions in smaller complexes,[26] those enzymes preferring $Mg^{2+}$ and $Ca^{2+}$[21] have predominantly oxygen ligands, such as the carboxyl groups of glutamate and aspartate, in the metal-binding site. The ions $Mg^{2-}$, $Mn^{2+}$, and $Ca^{2+}$ are relatively close to one another in their "hardness"[27–29] and also share similar binding elements. Although $Mn^{2+}$ is used in crystallographic structures (e.g., that for APE1) as a more electron-dense replacement for $Mg^{2+}$, $Mn^{2+}$ has a much wider variation in the type of contacts it makes with protein ligands. The sites in the enzymes studied here that preferentially use $Mn^{2+}$, such as the MreII nuclease[30] and the ser/thr protein phosphatase of phage λ[31] combine carboxyls, carboxyl oxygens of Asn or Gln, and imadazole nitrogens. The "softer" ions,

**TABLE V. Highest Scoring Proteins Found by MOTIFMINER in the ASTRAL40 Database Starting With PCP-Motifs Identified for an Alignment of IL-1 Related Proteins**

| PCPMer score | PDB ID | SCOP | Bound ion | Description |
|---|---|---|---|---|
| **1783** | **1L2H** | **b.42.1.2** | **No** | **Interleukin-1 beta (Hu)**[a] |
| **1682** | **1ILR** | **b.42.1.2** | **No** | **Interleukin-1 receptor antagonist protein (Hu)**[a] |
| **1658** | **2ILA** | **b.42.1.2** | **No** | **Interleukin-1 alpha (Hu)**[a] |
| 1638 | 1A28 | a.123.1.1 | No | Progesterone receptor (Hu) |
| 1627 | 1DL2 | a.102.2.1 | $Ca^{2+}$ | Class I alpha-1;2-mannosidase, catalytic domain (*S. cerevisiae*) |
| 1618 | 1IVY | c.69.1.5 | No | Human "protective protein," HPP (Hu) |
| 1611 | 1HLE | e.1.1.1 | No[b] | Elastase inhibitor (Horse) |
| 1610 | 1MQS | e.25.1.1 | No | Sly1P protein (*S. cerevisiae*)) |
| 1596 | 1AUI | d.159.1.3 | $Fe^{3+}$, $Zn^{2+}$ | Protein phosphatase-2B (calcineurin A subunit) (Hu) |
| 1595 | 1M7S | e.5.1.1 | heme | Catalase I (*Pseudomonas syringae*) |
| 1589 | 1DMU | c.52.1.4 | $Ca^{2+}$ | Restriction endonuclease BglI (*B. subtilis*) |
| 1588 | 1J5W | d.104.1.1 | No | Glycyl-tRNA synthetase (GlyRS) alpha chain (*Thermotoga maritima*, TM0216) |
| 1588 | 1C3P | c.42.1.2 | $Zn^{2+}$[c] | HDAC homologue (*Aquifex aeolicus*) |
| 1584 | 1AYM | b.10.1.4 | $Zn^{2+}$[d] | Rhinovirus coat protein (Hu rhinovirus 16) |
| 1578 | 1M0Z | c.10.2.7 | No | von Willebrand factor binding domain of glycoprotein Ib alpha (Hu) |
| **1569** | **1BFG** | **b.42.1.1** | **No** | **Basic FGF (FGF2) (Hu)**[a] |
| 1564 | 1CIP | c.37.1.8 | $Mg^{2+}$ | (A:32-60, A:182-347) Transducin (alpha subunit) (Rat) |
| 1559.50 | 1LL7 | c.1.8.5 | No | (A:36-292, A:355-427) Chitinase 1 (Fungus (*Coccidioides immitis*)) |
| 1550 | 1EU1 | c.81.1.1 | $Mo^{6+}$, $Cd^{2+}$ | Dimethylsulfoxide reductase (DMSO reductase) (*Rhodobacter sphaeroides*) |
| 1549 | 1MKF | b.116.1.1 | No | Viral chemokine binding protein m3 (Murine herpesvirus 4, Muhv-4) |
| 1540 | 1LST | c.94.1.1 | No | Lysine-, arginine-, ornithine-binding (LAO) protein (*Salmonella typhimurium*) |
| 1536 | 1QGI | d.2.1.7 | Sulfate | Endochitosanase (*Bacillus circulans*) |
| 1534 | 1B6C | d.144.1.1 | Sulfate | Type I TGF-beta receptor R4 (Hu) |
| 1530 | 1CXP | a.93.1.2 | Heme, $Ca^{2+}$ | Myeloperoxidase (Hu) |
| 1523 | 1FN9 | d.196.1.1 | $Zn^{2+}$ | Outer capsid protein sigma 3 (Reovirus) |
| 1519 | 1E6U | c.2.1.2 | $SO_4^{2-}$ | GDP-4-keto-6-deoxy-d-mannose epimerase/reductase (*E. coli*) |
| 1516 | 1GKY | c.37.1.1 | $SO_4^{2-}$, GMP | Guanylate kinase ((*S. cerevisiae*)) |
| 1497 | 1IOW | c.30.1.2 | $Mg^{2+}$ | D-Ala-D-Ala ligase, N-domain (1-96) (*E. coli*, gene dd1B) |
| 1492 | 2BPA | b.10.1.1 | No | Bacteriophage phi-X174 capsid proteins |
| 1491 | 1RUX | b.13.2.2 | No | Adenovirus hexon (Hu adenovirus type 5) |

[a] The four proteins in the initial DALI alignment used to define PCP-motifs are **bold.**
[b] Calcium ion identified in structure mediates a lattice contact.
[c] Not in crystal structure.
[d] Nonenzymatic $Zn^{2+}$ site between the subunits of the viral proteins on the surface of the virus.

$Zn^{2+}$ and $Fe^{2+}$,[28,32] can also be coordinated by nitrogens, which are presented by the molegos of both the dimetallic phosphatases, the dioxygenases and related VOCs. In sites where $Zn^{2+}$ plays a structural role, it is typically coordinated by cysteine and histidine residues.[33] However, no cysteine ligands are present in the active sites of the metalloenzymes of this study, and Zn is predominantly bound by carbonyl oxygens and histidine in these examples. The dimetallic phosphatase and dioxygenase boxes that are specific for $Fe^{2+}$ have conserved histidines in the binding positions, reflecting this ion's affinity for imadazole nitrogen.

Except for the DNase 1 family, the binding distances between the metal ion and the ligands are the same to within 0.1 Å in the crystal structures from each family. These results suggest that the basic architecture of the metal sites can be adapted to function by discrete sequence alterations that dictate metal ion specificity. In the dimetallic phosphatases, again regardless of metal ion bound, there is also a shared ligand between the two metals that are asymetrically bound. This shields the metals from one another and allows two metal ions to occupy about the same space that the single ones do in the other sites.

The metal ion in the two DNase 1 superfamily proteins available for analysis is more loosely bound to the residues in the active site than is the case for the other two families. This is consistent with a previous analysis for $Mg^{2+}$ binding, which indicated that this metal can accept up to only three negatively charged ligands, and fewer depending on the solvent accessibility of the binding site.[34] The only structure of these proteins that is consistent with the distances in the other two families is that containing $Pb^{2+}$, an element that does not support catalysis by APE1. Note that both the synaptojanin and APE1 structures have substrate bound, and the metal ion has ligands to the substrate in both cases (Table VII, supplementary information). MD simulations have suggested that the position of the metal ion differs in the free enzyme before and after cleavage of the substrate (Oezguen et al., forthcoming).

## DISCUSSION

This report shows that the PCPMer program can be used to analyze similar elements in the architecture of several families of metal-binding proteins, and distinguish homologues. The PCP-motifs defined for three distinct types of

**TABLE VI. Residues in the Metal-Binding Sites of the Proteins in This Study, as a Function of Metal Ion**

| Protein (PDB file name) | Metal | Binding site | | | | | |
|---|---|---|---|---|---|---|---|
| HuAPE1 (PDB structures for this enzyme bound to 3 different metal ions) | Mn, Sm, Pb[a] | PDB file: | | 1DE9<br>Mn | 1E9N | | 1BIX<br>Sm |
| | | | | | Pb1 | Pb2 | |
| | | ASP70 | OD1 | 4.13Å | | | 2.52Å |
| | | GLU96 | OE1 | | 1.99Å | | 2.75Å |
| | | GLU96 | OE2 | 2.42Å | | | 2.32Å |
| | | ASP210 | OD2 | 5.38Å | 2.79Å | | 6.8Å |
| | | ASN212 | ND2 | | | 2.98Å | |
| | | ASP308 | OD2 | 3.18Å | | | 4.4Å |
| Synaptojanin (1I9Z) | Ca | ASN568 | OD1 | 3.17 | | | |
| | | GLU597 | OE2 | 2.83 | | | |
| | | ASP838 | OD2 | 3.98 | | | |
| N5P (1USH) | 2Zn | | | ZN1 | ZN2 | | |
| | | ASP41 | OD2 | 2.06Å | | | |
| | | HIS43 | NE2 | 2.10Å | | | |
| | | ASP84 | OD2 | 2.34Å | 2.21Å[b] | | |
| | | ASN116 | OD1 | | 2.01Å | | |
| | | GLN254 | OE1 | 2.37Å | | | |
| | | HIS217 | NE2 | | 2.09Å | | |
| | | HIS252 | ND1 | | 2.21Å | | |
| Mre11 nuclease (1II7) | 2 Mn | | | MN403 | MN404 | | |
| | | ASP8 | OD1 | 2.12Å | | | |
| | | HIS10 | NE2 | 2.42Å | | | |
| | | ASP49 | OD1 | 2.26Å | 2.41Å[b] | | |
| | | ASN84 | OD1 | | 2.17Å | | |
| | | HIS173 | NE2 | | 2.22Å | | |
| | | HIS206 | ND1 | | 2.48Å | | |
| | | HIS208 | NE2 | 2.49Å | | | |
| Pig acid phosphatase (1UTE) | 2 Fe | | | FE1 | FE2 | | |
| | | ASP14 | OD2 | 2.11Å | | | |
| | | ASP52 | OD2 | 2.27Å | 2.40Å[b] | | |
| | | TYR55 | OH | 1.98Å | | | |
| | | HIS223 | NE2 | 2.32Å | | | |
| | | ASN91 | OD1 | | 2.24Å | | |
| | | HIS186 | NE2 | | 2.23Å | | |
| | | HIS221 | ND1 | | 2.37Å | | |
| Ser/Thr protein phosphatase (1G5B) | 2 Mn | | | MN1 | MN2 | | |
| | | ASP20 | OD2 | 2.39Å | | | |
| | | HIS22 | NE2 | 2.20Å | | | |
| | | ASP49 | OD2 | 2.22Å | 2.31Å[b] | | |
| | | ASN75 | OD1 | | 2.10Å | | |
| | | HIS139 | NE2 | | 2.18Å | | |
| | | HIS186 | ND1 | | 2.22Å | | |
| Catechol 2,3-dioxygenase (1MPY) | Fe | | | FE | | | |
| | | HIS153 | NE2 | 2.39Å | | | |
| | | HIS214 | NE2 | 2.50Å | | | |
| | | GLU265 | OE1 | 2.29Å | | | |
| 4-hydroxyphenyl pyruvate dioxygenase (1CJX) | Fe | | | FE | | | |
| | | HIS161 | NE2 | 2.18Å | | | |
| | | HIS240 | NE2 | 2.08Å | | | |
| | | GLU322 | OE1 | 1.96Å | | | |
| 2,3-dihydroxybiphenyl 1,2-dioxygenase (1HAN) | Fe | | | FE1 | FE2[c] | | |
| | | HIS146 | NE2 | 2.15Å | | | |
| | | HIS189 | NE2 | | 2.44Å[c] | | |
| | | HIS210 | NE2 | 2.25Å | | | |
| | | GLU260 | OE1 | 1.96Å | | | |
| Human glyoxylase (1QIP) | Zn | | | ZN | | | |
| | | GLN33 | OE1 | 2.03Å | | | |
| | | GLU99 | OE1 | 2.01Å | | | |
| | | HIS126 | NE2 | 2.03Å | | | |
| | | GLU172 | OE1 | 1.99Å | | | |

[a] APE1 has highest activity with Mg, but the 3 crystal structures have different metal ions. There are two Pb ions in the active site of IE9N.

[b] Bridging carbonyl between the two metal ions.

[c] The second Fe ion in the 1HAN structure is at the surface and probably has no effect on the active site.

metalloenzymes could be used in MOTIFMINER to identify the proteins in the initial alignment, as well as homologues with related functions. This indicates the PCPMer approach can aid in defining the function of proteins in genomic databases, when combined with other tools for identifying sequence similarity.

### Identifying Molego Architecture and Using PCPMer to Define Function

Assigning function to genomic sequences is a challenging problem that requires new approaches.[35–37] As blocks of homology become smaller, it becomes progressively more difficult to distinguish meaningful matches from random ones.[38–40] As these results show, when given a structure-based alignment of the sequences of homologous proteins, PCPMer can be used to detect similar metal ion binding proteins. For example, PCPMer scored a Ser/Thr protein phosphatase (1AUI) very highly as a dimetallic phosphatase even though it was not included in the initial alignment used to identify PCP-motifs in this family, and also listed several polymerases, which have two metals in their active centers, as being related to this family (Table II). The sensitivity of the approach is not limited to metal-containing proteins. When molegos were defined for the IL-1B family (Table V), the receptors for progesterone and TGF-β were identified, as well as a viral chemokine-binding protein. Our word-based approach can pinpoint underlying structural and functional similarities in proteins, regardless of the distance (and order) in the sequence between conserved elements. Thus, it is a particularly potent tool to identify areas of low but significant similarity in homologous proteins with overall low identity. As we previously showed, this is a very difficult task for other methods for determining sequence similarity.[4]

While segments of short sequence identity alone may indicate common structure,[41] it is generally accepted that both sequence and structural similarity is needed to establish functional homology.[42] For example, the DALI program,[13] which couples similar structure with sequence, is able to generate more meaningful alignments than programs, such as CLUSTALW,[43] which rely on sequence alone. Hence, we limited this study to protein families where several crystal structure representatives were known. These examples show that viewing individual elements as building blocks simplifies the analysis of residues that mediate specific metal binding. Our results indicate that PCPMer can be used to generate testable hypotheses about the function of novel proteins identified by genomic sequencing that are unclassified by conventional sequence analysis approaches.

The decomposition analysis of these proteins is particularly valuable when used to relate variations between proteins to substrate specificity and catalysis. Thus, it can play a useful role in protein design.[23,24,44]

### Metal Ion Specificity Despite Similarities in the Metal-Binding Mechanisms

One important result of our analysis was that while the molego structure, i.e., the protein architectural elements, of the active site within a family was relatively invariant, discrete changes in a few key residues may dictate the metal ion specificity for catalysis. As Tables VI and Vll (supplementary information) indicate, the conserved amino acid positions alter with the metal type, within the limitations of the site geometry (the actual occupancy will of course also be affected by the relative concentrations of the metals in the biological environment). All three metal-binding sites have a key carboxylate linkage, and then other ligands that vary with the preferentially bound metal ion. The exact pattern of variation will require more comparisons of sequences from enzymes in the families. This result suggests that once a metal ion–binding site is defined, simple residue changes at defined positions can be made to alter its metal ion specificity.

The present speed of the PCPMer program is now sufficient to use it to scan for functional homologues in larger, genomic sequence databases (Bin Zhou et al., forthcoming). We are also testing more automatic approaches to PCP-motif generation, using for example a molego library assembled from existing data, such as that in PFAM.[45,46] Structural comparisons of the individual elements from many proteins, as described here, should establish a basic protein dictionary of the amino acid words that make up complex proteins.

## CONCLUSIONS

1. The metal containing active sites of three distinct enzyme groups, DNase 1 homologues, dimetallic phosphatases, and dioxygenases, can be decomposed into molegos, areas of conserved sequence and structure. The dimensions of the site and orientation of the molegos to the metal ions vary little across the superfamily members, even in homologues that have quite different overall activity.
2. The PCPMer program can be used to mine sequence databases and identify proteins with functional and structural similarities to a given protein family.
3. The residues in the binding site created by the molegos dictate the specificity for the type of metal ion bound by the metalloenzyme. The specific residue interactions with the metal ion observed in the enzymes in this study are consistent with rules established by previous biophysical studies of metal ion binding affinities.

## REFERENCES

1. Black CB, Huang HW, Cowan JA. Biological coordination chemistry of magnesium, sodium, and potassium ions. Protein and nucleotide binding sites. Coord Chem Rev 1994;135:165–202.
2. Engelman A, Craigie R. Efficient Magnesium-dependent human immunodeficiency virus type 1 intergrase activity. J Virol 1995;69:5908–5911.
3. Misra VK, Draper DE. A thermodynamic framework for Mg2+ binding to RNA. Proc Natl Acad Sci USA 2001;98:12456–12461.
4. Mathura VS, Schein CH, Braun W. Identifying property based sequence motifs in protein families and superfamilies: application

to DNase I related repair endonuclease. Bioinformatics 2003;19: 1381–1390.

5. Schein CH, Özgün N, Izumi T, Braun W. Total sequence decomposition distinguishes functional modules, "molegos" in apurinic/apyrimidinic endonucleases. BMC-Bioinformatics 2002;3:37.
6. Gilardi G, Fantuzzi A, Sadeghi S. Engineering and design in the bioelectrochemistry of metalloproteins. Curr Opin Struct Biol 2001;11:491–499.
7. Gilardi G, Meharenna Y, Tsotsou G, Sadeghi S, Fairhead M, Giannini S. Molecular lego: design of molecular assemblies of P450 enzymes for nanobiotechnology. Biosens Bioelectron 2002;17:133–145.
8. Venkatarajan MS, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. J Mol Model 2001;7:445–453.
9. Izumi T, Schein CH, Oezguen N, Feng Y, Braun W. Effects of backbone contacts 3′ to the abasic site on the cleavage and the product binding by human apurinic/apyrimidinic endonuclease (APE1). Biochemistry 2004;43:684–689.
10. Serre L, Sailland A, Sy D, Boudec P, Rolland A, Pebay-Peyroula E, Cohen-Addad C. Crystal structure of Pseudomonas fluorescens 4-hydroxyphenylpyruvate dioxygenase: an enzyme involved in the tyrosine degradation pathway. Structure Fold Des 1999;7:977–988.
11. Gerlt JA, Babbitt PC. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Ann Rev Biochem 2001;70:209–246.
12. Zhu H, Schein CH, Braun W. MASIA: recognition of common patterns and properties in multiple aligned protein sequences. Bioinformatics 2000;16:950–951.
13. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26:316–319.
14. Knöfel T, Sträter N. X-ray structure of the *Escherichia coli* periplasmic 5′-nucleotidase containing a dimetal catalytic site. Nature Struct Biol 1999;6:448–453.
15. Knöfel T, Sträter N. Mechanism of hydrolysis of phosphate esters by the dimetal center of 5′-nucleotidase based on crystal structures. J Mol Biol 2001;309:239–254.
16. Armstrong RN. Mechanistic diversity in a metalloenzyme superfamily. Biochemistry 2000;39:13625–13632.
17. Schein CH. The shape of the messenger: using protein structure information to design novel cytokine-based therapeutics. Curr Pharmaceut Des 2002;8:2113–2129.
18. Chandonia JM, Walker NS, Conte LL, Koehl P, Levitt M, Brenner SE. ASTRAL compendium enhancements. Nucleic Acids Res 2002;30:260–263.
19. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for sequence and structure analysis. Nucleic Acids Res 2000;28:254–256.
20. Koradi R, Billeter M, Wüthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 1996;14:51–55.
21. Katz AK, Glusker JP, Beebe SA, Bock CW. Calcium ion coordination: a comparison with that of beryllium, magnesium, and zinc. J Am Chem Soc 1996;118:5752–5763.
22. Hymowitz SG, O'Connell MP, Ultsch MH, Hurst A, Totpal K, Ashkenazi A, de Vos AM, Kelley RF. A unique zinc-binding site revealed by a high resolution X-ray structure of homotrimeric Apo2L/TRAIL. Biochemistry 2000;39:633–640.
23. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A. De novo design and structural characterization of proteins and metalloproteins. Annu Rev Biochem 1999;68:779–819.
24. Dwyer MA, Looger LL, Hellinga HW. Computational design of a Zn2+ receptor that controls bacterial gene expression. Proc Natl Acad Sci USA 2003;100:11255–11260.
25. Trachtman M, Markham GD, Glusker JP, George P, Bock CW. Interactions of metal ions with water: ab initio molecular orbital studies of structure, vibrational frequencies, charge distributions, bonding enthalpies, and deprotonation enthalpies. 2. Monohydroxides. Inorg Chem 2001;40:4230–4241.
26. Bock CW, Katz AK, Markham GD, Glusker JP. Manganese as a replacement for magnesium and zinc: functional comparison of the divalent ions. J Am Chem Soc 1999;121:7360–7372.
27. Galigniana MD, Piwien-Pilipuk G. Comparative inhibition by hard and soft metal ions of steroid-binding capacity of renal mineralcorticoid receptor cross-linked to the 90-kDa heat-shock protein heterocomplex. Biochem J 1999;341:585–592.
28. Glassman T, Klopman, G., Cooper C. Use of the generalized perturbation theory to predict the interaction of purine nucleotides with metal ions. Biochemistry 1973;12:5013–5019.
29. Klopman G. Chemical reactivity and the concept of charge- and frontier-controlled reactions. J Am Chem Soc 1968;90:223–234.
30. Hopfner K-P, Karcher A, Craig L, Woo TT, Carney JP, Tainer JA. Structural biochemistry and interaction architecture of the DNA double-strand break repair Mre11 nuclease and Rad50-ATPase. Cell 2001;105:473–485.
31. Voegtli WC, White DJ, Reiter NJ, Rusnak F, Rosenzweig AC. Structure of the bacteriophage lambda Ser/Thr protein phosphatase with sulfate ion bound in two coordination modes. Biochemistry 2000;39:15365–15374.
32. Dudev T, Lim C. Metal Selectivity in metalloproteins: $Zn^{2+}$ vs. $Mg^{2+}$. J Phys Chem B 2001;105:4446–4452.
33. Maret W. Exploring the zinc proteome. J Analyt Atomic Spec 2004;19:15–19.
34. Dudev T, Lim C. Principles governing Mg, Ca, and Zn binding and selectivity in proteins. Chem Rev 2003;103:773–787.
35. Ben-Hur A, Brutlag D. Remote homology detection: a motif based approach. Bioinformatics 2003;19(Suppl. 1):i26–i33.
36. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 2003;21:697– 700.
37. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomies. Nat Struct Biol 2001;8:559–566.
38. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
39. Abagyan RA, Batalov S, Cardozo T, Totrov M, Webber J, Zhou YY. Homology modeling with internal coordinate mechanics: deformation zone mapping and improvements of models via conformational search. Proteins 1997; (Suppl. 1):29–37.
40. Mehta PK, Argos P, Barbour AD, Christen P. Recognizing very distant sequence relationships among proteins by family profile analysis. Proteins 1999;35:387–400.
41. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASPIII targets using ROSETTA. Proteins 1999;(Suppl. 3):171–176.
42. Lichtarge O, Yamamoto KR, Cohen FE. Identification of functional surfaces of the zinc binding domains of intracellular receptors. J Mol Biol 1997;274:325–337.
43. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position- specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673– 4680.
44. Summa C, Rosenblatt M, Hong J-K, Lear J, DeGrado W. Computational de novo Design, and Characterization of an A2B2 Diiron Protein. J Mol Biol 2002;321:923–938.
45. Andrade MA, Casari G, Sander C, Valencia A. Classification of protein families and detection of the determinant residues with an improved self-organizing map. Biol Cybernet 1997;76:441–450.
46. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Prot Sci 1998;7:2469–2471.